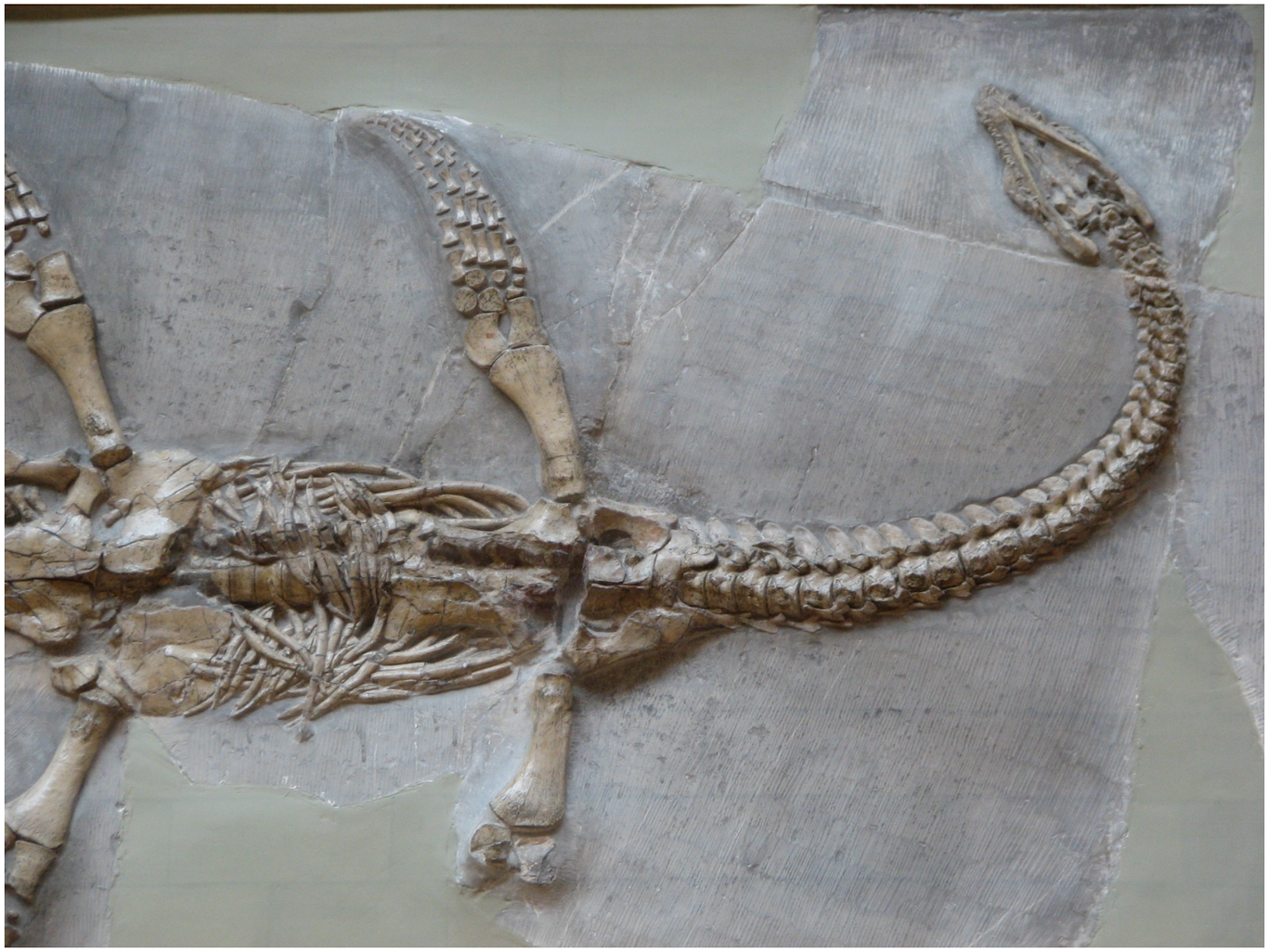
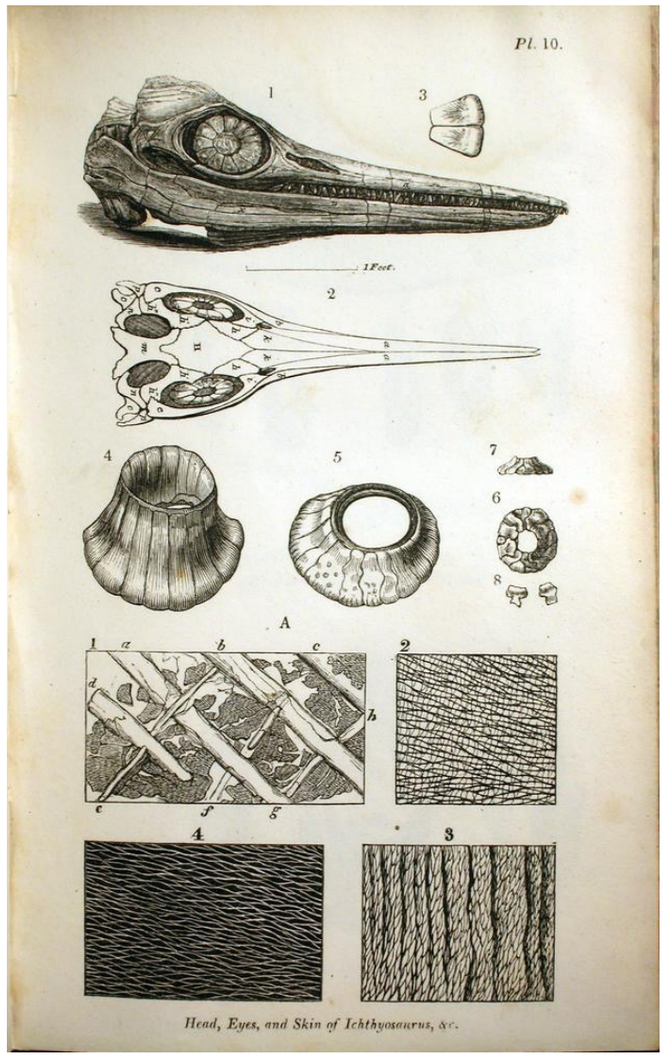
GeoDeepDive: A Cyberinfrastructure to Support Text and Data Mining

Ian Ross, Miron Livny, Tim Theisen
Center for High Throughput Computing
University of Wisconsin-Madison USA

Shanan E. Peters, John Czaplewski
Department of Geoscience
University of Wisconsin-Madison USA

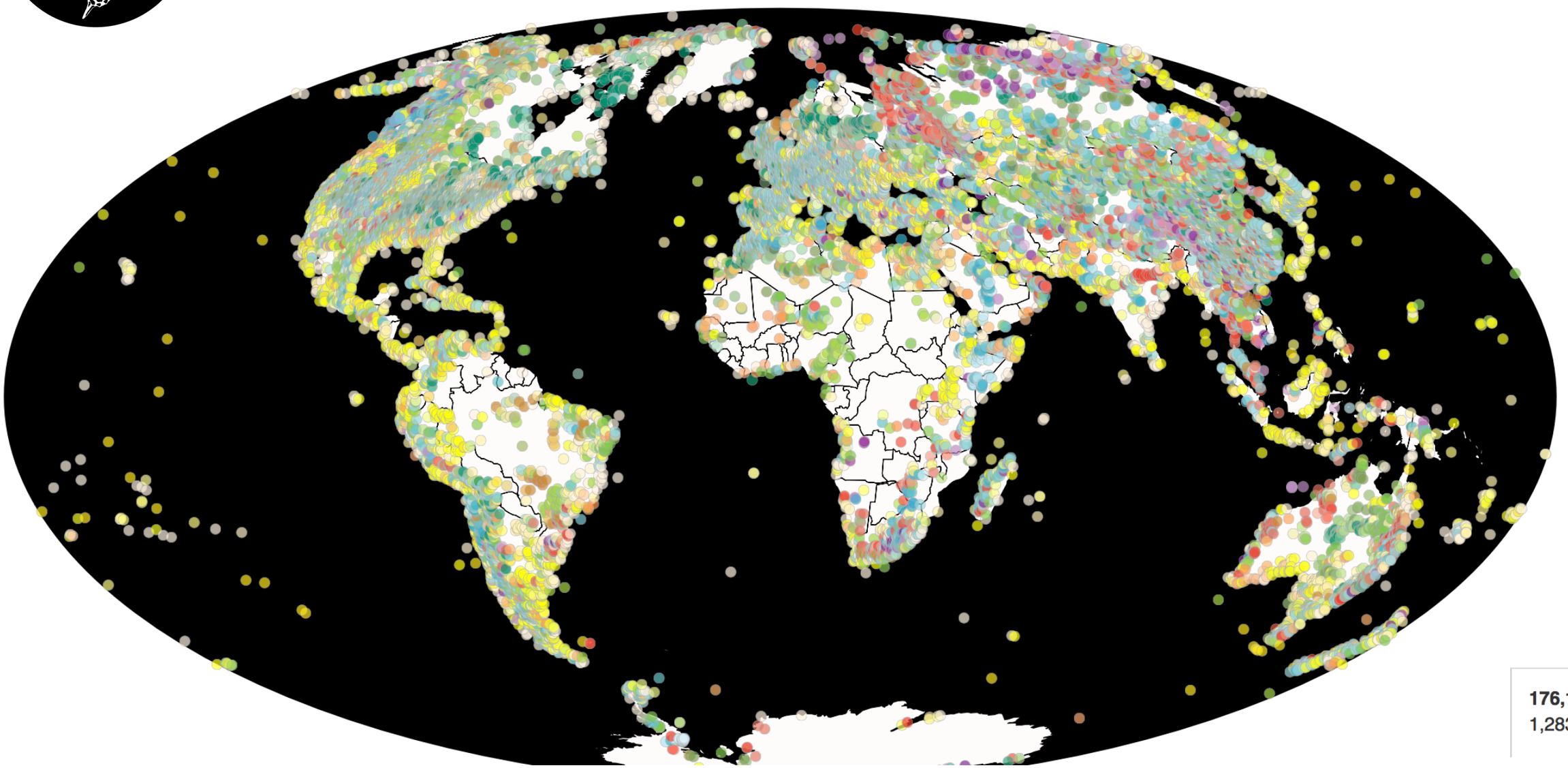
How has biodiversity changed on Earth over geologic time?





Paleobiology Database

~10 continuous person years of effort



Mollusca 33%

Brachiopoda 13%

Arthropoda 10%

Mammalia 8%

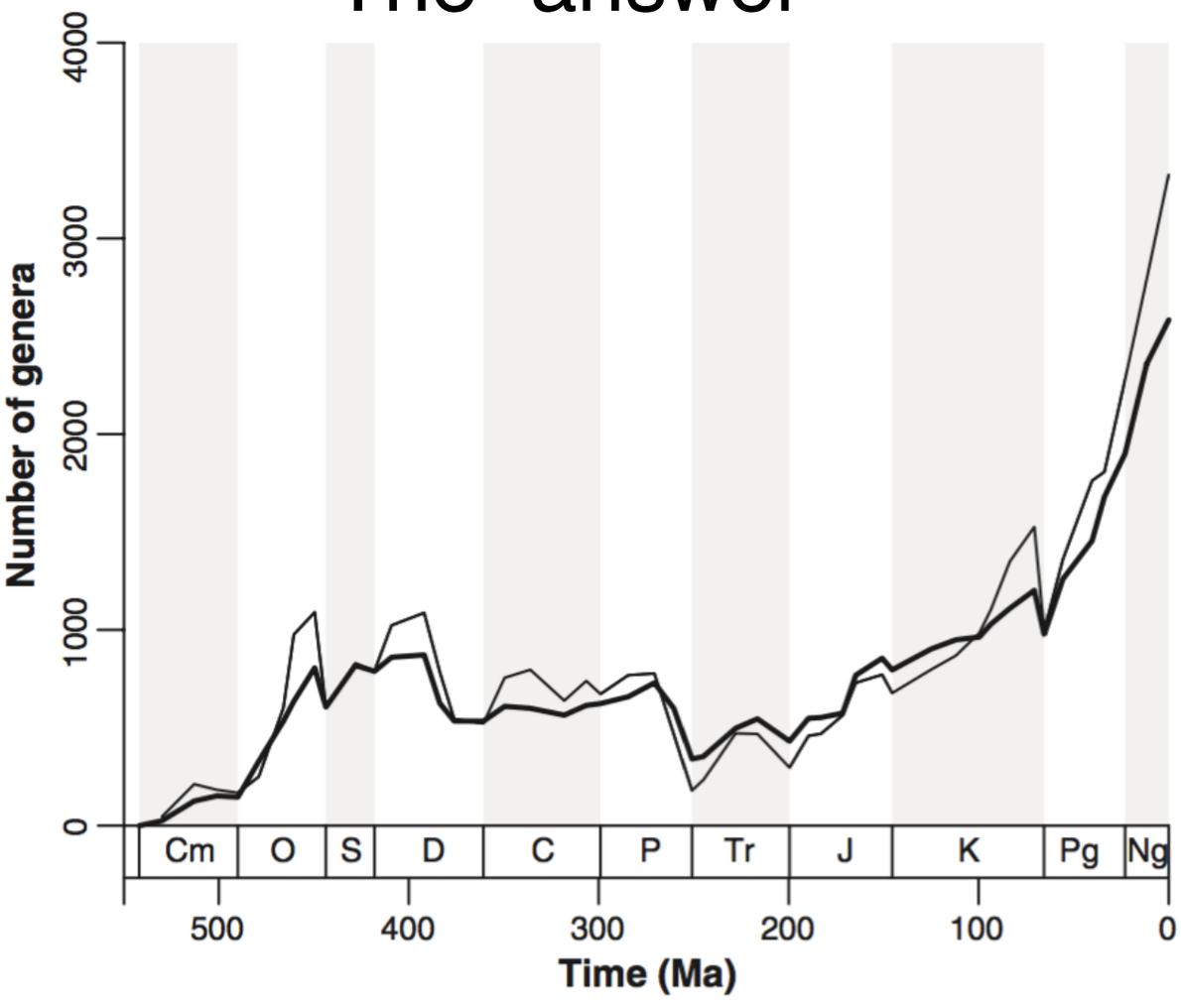
Cnidaria 5%

Radiolaria 4%

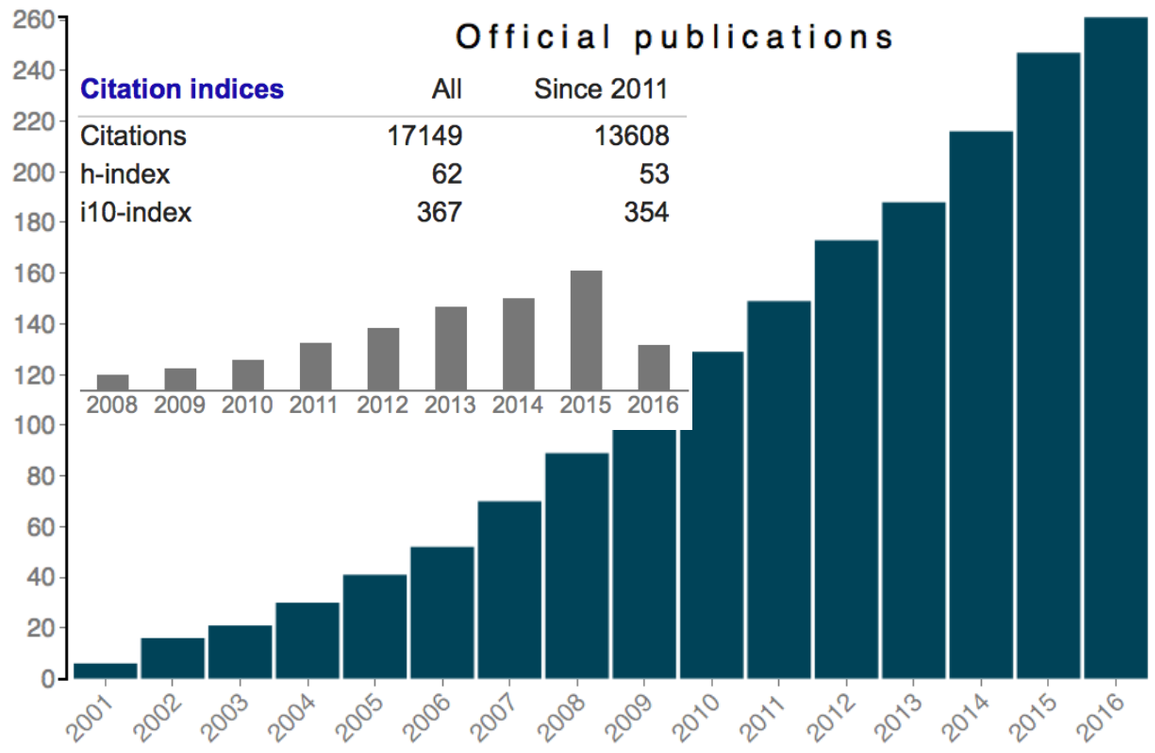
176,170 total collections
1,283,450 total occurrence

How has biodiversity changed on Earth over geologic time?

The “answer”



The combined effort has paid off!



Can a machine reading system reproduce the PBDB?

machine

DeepDive v0.8.0

Think about **features**, not **algorithms**.

http://deepdive.stanford.edu

vs.

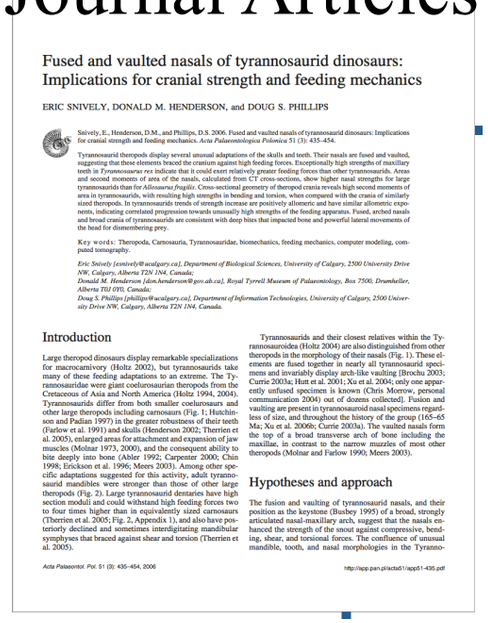
humans

The Paleobiology Database

revealing the history of life

“Data entry” = feature identification and extraction

Journal Articles



Fused and vaulted nasals of tyrannosaurid dinosaurs: Implications for cranial strength and feeding mechanics

ERIC SNIVELY, DONALD M. HENDERSON, and DOUG S. PHILLIPS

Snively, E., Henderson, D.M., and Phillips, D.S. 2006. Fused and vaulted nasals of tyrannosaurid dinosaurs: Implications for cranial strength and feeding mechanics. *Acta Palaeontologica Polonica* 51 (3): 435–454.

Tyrannosaurid theropods display several unusual adaptations of the skulls and teeth. Their nasals are fused and vaulted, suggesting that these elements braced the cranium against high feeding forces. Exceptionally high strengths of maxillary teeth in *Tyrannosaurus rex* indicate that it could exert relatively greater feeding forces than other tyrannosaurids. Areas and second moments of area of the nasals, calculated from CT cross-sections, show higher nasal strengths for large tyrannosaurids than for *Allosaurus fragilis*. Cross-sectional geometry of theropod crania reveals high second moments of area in tyrannosaurids, with resulting high strengths in bending and torsion, when compared with the crania of similarly sized theropods. In tyrannosaurids trends of strength increase are positively allometric and have similar allometric exponents, indicating correlated progression towards unusually high strengths of the feeding apparatus. Fused, arched nasals and broad crania of tyrannosaurids are consistent with deep bites that impacted bone and powerful lateral movements of the head for dismembering prey.

Key words: Theropoda, Carnosauria, Tyrannosauridae, biomechanics, feeding mechanics, computer modeling, computed tomography.

Eric Snively [esnively@ucalgary.ca], Department of Biological Sciences, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada;
Donald M. Henderson [don.henderson@gov.ab.ca], Royal Tyrrell Museum of Palaeontology, Box 7500, Drumheller, Alberta T0J 0Y0, Canada;
Doug S. Phillips [phillips@ucalgary.ca], Department of Information Technologies, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada.

Introduction

Large theropod dinosaurs display remarkable specializations for macrocarnivory (Holtz 2002), but tyrannosaurids take many of these feeding adaptations to an extreme. The Tyrannosauridae were giant coelurosaurian theropods from the Cretaceous of Asia and North America (Holtz 1994, 2004). Tyrannosaurids differ from both smaller coelurosaurs and other large theropods including carnosaurs (Fig. 1; Hutchinson and Padian 1997) in the greater robustness of their teeth (Farlow et al. 1991) and skulls (Henderson 2002; Therrien et al. 2005), enlarged areas for attachment and expansion of jaw muscles (Molnar 1973, 2000), and the consequent ability to bite deeply into bone (Abler 1992; Carpenter 2000; Chin 1998; Erickson et al. 1996; Meers 2003). Among other specific adaptations suggested for this activity, adult tyrannosaurid mandibles were stronger than those of other large theropods (Fig. 2). Large tyrannosaurid dentaries have high section moduli and could withstand high feeding forces two to four times higher than in equivalently sized carnosaurs (Therrien et al. 2005; Fig. 2, Appendix 1), and also have posteriorly declined and sometimes interdigitating mandibular symphyses that braced against shear and torsion (Therrien et al. 2005).

Tyrannosaurids and their closest relatives within the Tyrannosauroidea (Holtz 2004) are also distinguished from other theropods in the morphology of their nasals (Fig. 1). These elements are fused together in nearly all tyrannosaurid specimens and invariably display arch-like vaulting [Brochu 2003; Currie 2003a; Hutt et al. 2001; Xu et al. 2004; only one apparently unfused specimen is known (Chris Morrow, personal communication 2004) out of dozens collected]. Fusion and vaulting are present in tyrannosaurid nasal specimens regardless of size, and throughout the history of the group (165–65 Ma; Xu et al. 2006b; Currie 2003a). The vaulted nasals form the top of a broad transverse arch of bone including the maxillae, in contrast to the narrow muzzles of most other theropods (Molnar and Farlow 1990; Meers 2003).

Hypotheses and approach

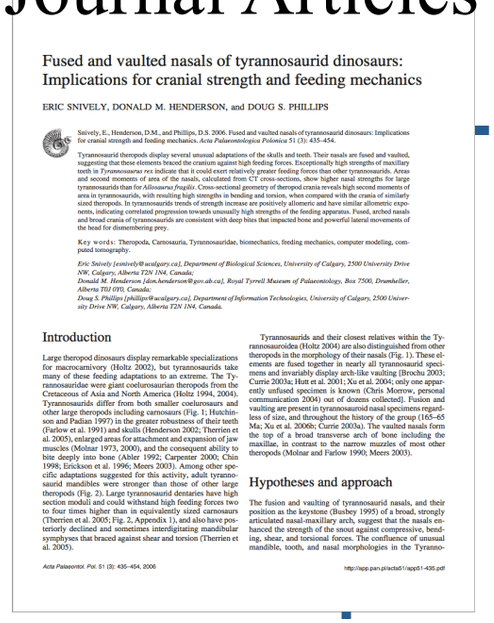
The fusion and vaulting of tyrannosaurid nasals, and their position as the keystone (Busbey 1995) of a broad, strongly articulated nasal-maxillary arch, suggest that the nasals enhanced the strength of the snout against compressive, bending, shear, and torsional forces. The confluence of unusual mandible, tooth, and nasal morphologies in the Tyranno-

Acta Palaeontol. Pol. 51 (3): 435–454, 2006

http://app.pan.pl/acta51/app51-435.pdf

“Data entry” = feature identification and extraction

Journal Articles



Fused and vaulted nasals of tyrannosaurid dinosaurs: Implications for cranial strength and feeding mechanics

ERIC SNIVELY, DONALD M. HENDERSON, and DOUG S. PHILLIPS

Snively, E., Henderson, D.M., and Phillips, D.S. 2006. Fused and vaulted nasals of tyrannosaurid dinosaurs: Implications for cranial strength and feeding mechanics. *Acta Palaeontologica Polonica* 51 (3): 435–454.

Tyrannosaurid theropods display several unusual adaptations of the skulls and teeth. Their nasals are fused and vaulted, suggesting that these elements braced the cranium against high feeding forces. Exceptionally high strengths of maxillary teeth in *Tyrannosaurus rex* indicate that it could exert relatively greater feeding forces than other tyrannosaurids. Areas and second moments of area of the nasals, calculated from CT cross-sections, show higher nasal strengths for large tyrannosaurids than for *Allosaurus fragilis*. Cross-sectional geometry of theropod crania reveals high second moments of area in tyrannosaurids, with resulting high strengths in bending and torsion, when compared with the crania of similarly sized theropods. In tyrannosaurids trends of strength increase are positively allometric and have similar allometric exponents, indicating correlated progression towards unusually high strengths of the feeding apparatus. Fused, arched nasals and broad crania of tyrannosaurids are consistent with deep bites that impacted bone and powerful lateral movements of the head for dismembering prey.

Key words: Theropoda, Carnosauria, Tyrannosauridae, biomechanics, feeding mechanics, computer modeling, computed tomography.

Eric Snively [esnively@ucalgary.ca], Department of Biological Sciences, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada;
Donald M. Henderson [don.henderson@gov.ab.ca], Royal Tyrrell Museum of Palaeontology, Box 7500, Drumheller, Alberta T0J 0Y0, Canada;
Doug S. Phillips [phillips@ucalgary.ca], Department of Information Technologies, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada.

Introduction

Large theropod dinosaurs display remarkable specializations for macrocarnivory (Holtz 2002), but tyrannosaurids take many of these feeding adaptations to an extreme. The Tyrannosauridae were giant coelurosaurian theropods from the Cretaceous of Asia and North America (Holtz 1994, 2004). Tyrannosaurids differ from both smaller coelurosaurs and other large theropods including carnosaurs (Fig. 1; Hutchinson and Padian 1997) in the greater robustness of their teeth (Farlow et al. 1991) and skulls (Henderson 2002; Therrien et al. 2005), enlarged areas for attachment and expansion of jaw muscles (Molnar 1973, 2000), and the consequent ability to bite deeply into bone (Abler 1992; Carpenter 2000; Chin 1998; Erickson et al. 1996; Meers 2003). Among other specific adaptations suggested for this activity, adult tyrannosaurid mandibles were stronger than those of other large theropods (Fig. 2). Large tyrannosaurid dentaries have high section moduli and could withstand high feeding forces two to four times higher than in equivalently sized carnosaurs (Therrien et al. 2005; Fig. 2, Appendix 1), and also have posteriorly declined and sometimes interdigitating mandibular symphyses that braced against shear and torsion (Therrien et al. 2005).

Tyrannosaurids and their closest relatives within the Tyrannosauroidea (Holtz 2004) are also distinguished from other theropods in the morphology of their nasals (Fig. 1). These elements are fused together in nearly all tyrannosaurid specimens and invariably display arch-like vaulting [Brochu 2003; Currie 2003a; Hutt et al. 2001; Xu et al. 2004; only one apparently unfused specimen is known (Chris Morrow, personal communication 2004) out of dozens collected]. Fusion and vaulting are present in tyrannosaurid nasal specimens regardless of size, and throughout the history of the group (165–65 Ma; Xu et al. 2006b; Currie 2003a). The vaulted nasals form the top of a broad transverse arch of bone including the maxillae, in contrast to the narrow muzzles of most other theropods (Molnar and Farlow 1990; Meers 2003).

Hypotheses and approach

The fusion and vaulting of tyrannosaurid nasals, and their position as the keystone (Busbey 1995) of a broad, strongly articulated nasal-maxillary arch, suggest that the nasals enhanced the strength of the snout against compressive, bending, shear, and torsional forces. The confluence of unusual mandible, tooth, and nasal morphologies in the Tyranno-

Acta Palaeontol. Pol. 51 (3): 435–454, 2006

http://app.pan.pl/acta51/app51-435.pdf

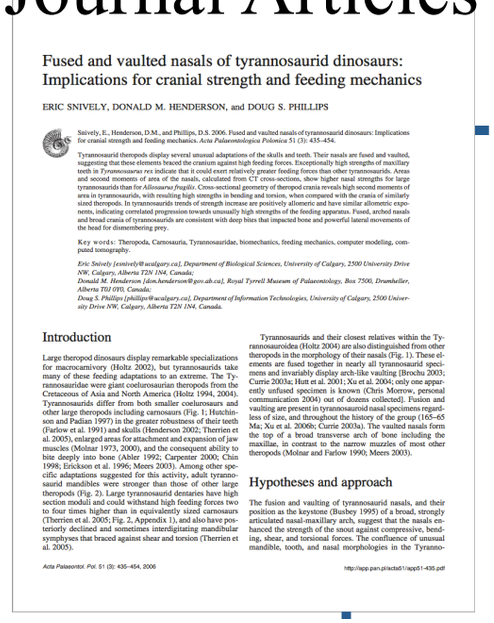
OCR

Text

> … *The Namurian Tsingyuan Formation from Ningxia, China, is divided into three members…*

“Data entry” = feature identification and extraction

Journal Articles



Fused and vaulted nasals of tyrannosaurid dinosaurs: Implications for cranial strength and feeding mechanics

ERIC SNIVELY, DONALD M. HENDERSON, and DOUG S. PHILLIPS

Snively, E., Henderson, D.M., and Phillips, D.S. 2006. Fused and vaulted nasals of tyrannosaurid dinosaurs: Implications for cranial strength and feeding mechanics. *Acta Palaeontologica Polonica* 51 (3): 435–454.

Tyrannosaurid theropods display several unusual adaptations of the skulls and teeth. Their nasals are fused and vaulted, suggesting that these elements braced the cranium against high feeding forces. Exceptionally high strengths of maxillary teeth in *Tyrannosaurus rex* indicate that it could exert relatively greater feeding forces than other tyrannosaurids. Areas and second moments of area of the nasals, calculated from CT cross-sections, show higher nasal strengths for large tyrannosaurids than for *Allosaurus fragilis*. Cross-sectional geometry of theropod crania reveals high second moments of area in tyrannosaurids, with resulting high strengths in bending and torsion, when compared with the crania of similarly sized theropods. In tyrannosaurids trends of strength increase are positively allometric and have similar allometric exponents, indicating correlated progression towards unusually high strengths of the feeding apparatus. Fused, arched nasals and broad crania of tyrannosaurids are consistent with deep bites that impacted bone and powerful lateral movements of the head for dismembering prey.

Key words: Theropoda, Carnosauria, Tyrannosauridae, biomechanics, feeding mechanics, computer modeling, computed tomography.

Eric Snively [esnively@ucalgary.ca], Department of Biological Sciences, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada;
Donald M. Henderson [don.henderson@gov.ab.ca], Royal Tyrrell Museum of Palaeontology, Box 7500, Drumheller, Alberta T0J 0Y0, Canada;
Doug S. Phillips [phillips@ucalgary.ca], Department of Information Technologies, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada.

Introduction

Large theropod dinosaurs display remarkable specializations for macrocarnivory (Holtz 2002), but tyrannosaurids take many of these feeding adaptations to an extreme. The Tyrannosauridae were giant coelurosaurian theropods from the Cretaceous of Asia and North America (Holtz 1994, 2004). Tyrannosaurids differ from both smaller coelurosaurs and other large theropods including carnosaurs (Fig. 1; Hutchinson and Padian 1997) in the greater robustness of their teeth (Farlow et al. 1991) and skulls (Henderson 2002; Therrien et al. 2005), enlarged areas for attachment and expansion of jaw muscles (Molnar 1973, 2000), and the consequent ability to bite deeply into bone (Abler 1992; Carpenter 2000; Chin 1998; Erickson et al. 1996; Meers 2003). Among other specific adaptations suggested for this activity, adult tyrannosaurid mandibles were stronger than those of other large theropods (Fig. 2). Large tyrannosaurid dentaries have high section moduli and could withstand high feeding forces two to four times higher than in equivalently sized carnosaurs (Therrien et al. 2005; Fig. 2, Appendix 1), and also have posteriorly declined and sometimes interdigitating mandibular symphyses that braced against shear and torsion (Therrien et al. 2005).

Tyrannosaurids and their closest relatives within the Tyrannosauroidea (Holtz 2004) are also distinguished from other theropods in the morphology of their nasals (Fig. 1). These elements are fused together in nearly all tyrannosaurid specimens and invariably display arch-like vaulting [Brochu 2003; Currie 2003a; Hutt et al. 2001; Xu et al. 2004; only one apparently unfused specimen is known (Chris Morrow, personal communication 2004) out of dozens collected]. Fusion and vaulting are present in tyrannosaurid nasal specimens regardless of size, and throughout the history of the group (165–65 Ma; Xu et al. 2006b; Currie 2003a). The vaulted nasals form the top of a broad transverse arch of bone including the maxillae, in contrast to the narrow muzzles of most other theropods (Molnar and Farlow 1990; Meers 2003).

Hypotheses and approach

The fusion and vaulting of tyrannosaurid nasals, and their position as the keystone (Busbey 1995) of a broad, strongly articulated nasal-maxillary arch, suggest that the nasals enhanced the strength of the snout against compressive, bending, shear, and torsional forces. The confluence of unusual mandible, tooth, and nasal morphologies in the Tyranno-

Acta Palaeontol. Pol. 51 (3): 435–454, 2006

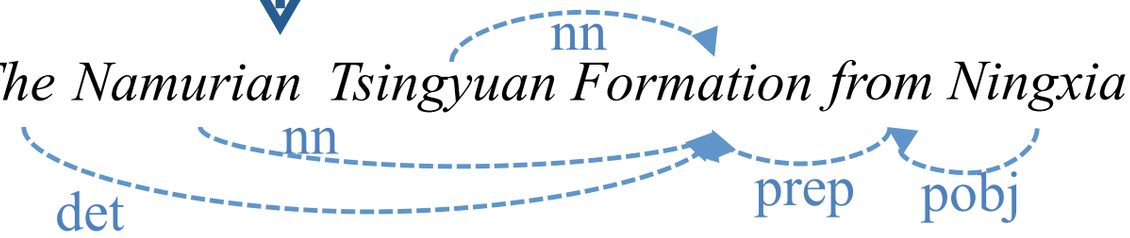
http://app.pan.pl/acta51/app51-435.pdf

OCR

Text

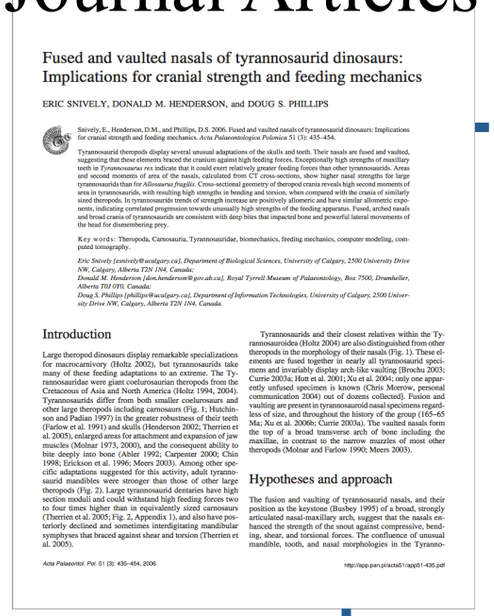
… *The Namurian Tsingyuan Formation from Ningxia, China, is divided into three members…*

NLP



“Data entry” = feature identification and extraction

Journal Articles



Fused and vaulted nasals of tyrannosaurid dinosaurs: Implications for cranial strength and feeding mechanics

ERIC SNIVELY, DONALD M. HENDERSON, and DOUG S. PHILLIPS

Snively, E., Henderson, D.M., and Phillips, D.S. 2006. Fused and vaulted nasals of tyrannosaurid dinosaurs: Implications for cranial strength and feeding mechanics. *Acta Palaeontologica Polonica* 51 (3): 435–454.

Tyrannosaurid theropods display several unusual adaptations of the skulls and teeth. Their nasals are fused and vaulted, suggesting that these elements braced the cranium against high feeding forces. Exceptionally high strengths of maxillary teeth in *Tyrannosaurus rex* indicate that it could exert relatively greater feeding forces than other tyrannosaurids. Areas and second moments of area of the nasals, calculated from CT cross-sections, show higher nasal strengths for large tyrannosaurids than for *Allosaurus fragilis*. Cross-sectional geometry of theropod crania reveals high second moments of area in tyrannosaurids, with resulting high strengths in bending and torsion, when compared with the crania of similarly sized theropods. In tyrannosaurids trends of strength increase are positively allometric and have similar allometric exponents, indicating correlated progression towards unusually high strengths of the feeding apparatus. Fused, arched nasals and broad crania of tyrannosaurids are consistent with deep bites that impacted bone and powerful lateral movements of the head for dismembering prey.

Key words: Theropoda, Carnosauria, Tyrannosauridae, biomechanics, feeding mechanics, computer modeling, computed tomography.

Eric Snively [esnively@ucalgary.ca], Department of Biological Sciences, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada;
Donald M. Henderson [don.henderson@gov.ab.ca], Royal Tyrrell Museum of Palaeontology, Box 7500, Drumheller, Alberta T0J 0Y0, Canada;
Doug S. Phillips [phillips@ucalgary.ca], Department of Information Technologies, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada.

Introduction

Large theropod dinosaurs display remarkable specializations for macrocarnivory (Holtz 2002), but tyrannosaurids take many of these feeding adaptations to an extreme. The Tyrannosauridae were giant coelurosaurian theropods from the Cretaceous of Asia and North America (Holtz 1994, 2004). Tyrannosaurids differ from both smaller coelurosaurs and other large theropods including carnosaurs (Fig. 1; Hutchinson and Padian 1997) in the greater robustness of their teeth (Farlow et al. 1991) and skulls (Henderson 2002; Therrien et al. 2005), enlarged areas for attachment and expansion of jaw muscles (Molnar 1973, 2000), and the consequent ability to bite deeply into bone (Abler 1992; Carpenter 2000; Chin 1998; Erickson et al. 1996; Meers 2003). Among other specific adaptations suggested for this activity, adult tyrannosaurid mandibles were stronger than those of other large theropods (Fig. 2). Large tyrannosaurid dentaries have high section moduli and could withstand high feeding forces two to four times higher than in equivalently sized carnosaurs (Therrien et al. 2005; Fig. 2, Appendix 1), and also have posteriorly declined and sometimes interdigitating mandibular symphyses that braced against shear and torsion (Therrien et al. 2005).

Tyrannosaurids and their closest relatives within the Tyrannosauroidea (Holtz 2004) are also distinguished from other theropods in the morphology of their nasals (Fig. 1). These elements are fused together in nearly all tyrannosaurid specimens and invariably display arch-like vaulting [Brochu 2003; Currie 2003a; Hutt et al. 2001; Xu et al. 2004; only one apparently unfused specimen is known (Chris Morrow, personal communication 2004) out of dozens collected]. Fusion and vaulting are present in tyrannosaurid nasal specimens regardless of size, and throughout the history of the group (165–65 Ma; Xu et al. 2006b; Currie 2003a). The vaulted nasals form the top of a broad transverse arch of bone including the maxillae, in contrast to the narrow muzzles of most other theropods (Molnar and Farlow 1990; Meers 2003).

Hypotheses and approach

The fusion and vaulting of tyrannosaurid nasals, and their position as the keystone (Busbey 1995) of a broad, strongly articulated nasal-maxillary arch, suggest that the nasals enhanced the strength of the snout against compressive, bending, shear, and torsional forces. The confluence of unusual mandible, tooth, and nasal morphologies in the Tyranno-

Acta Palaeontol. Pol. 51 (3): 435–454, 2006 http://app.pan.pl/acta51/app51-435.pdf

OCR

Text

… *The Namurian Tsingyuan Formation from Ningxia, China, is divided into three members…*

NLP

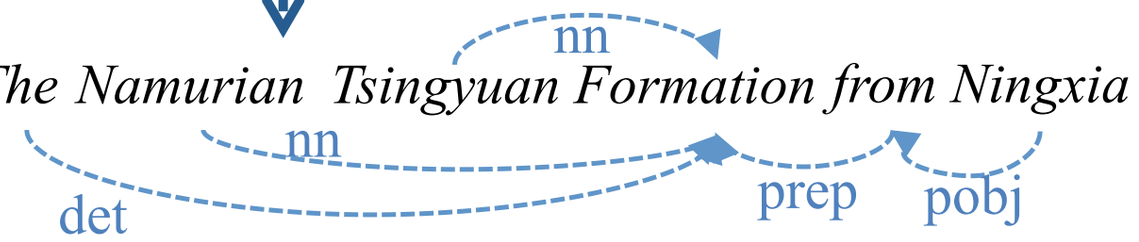


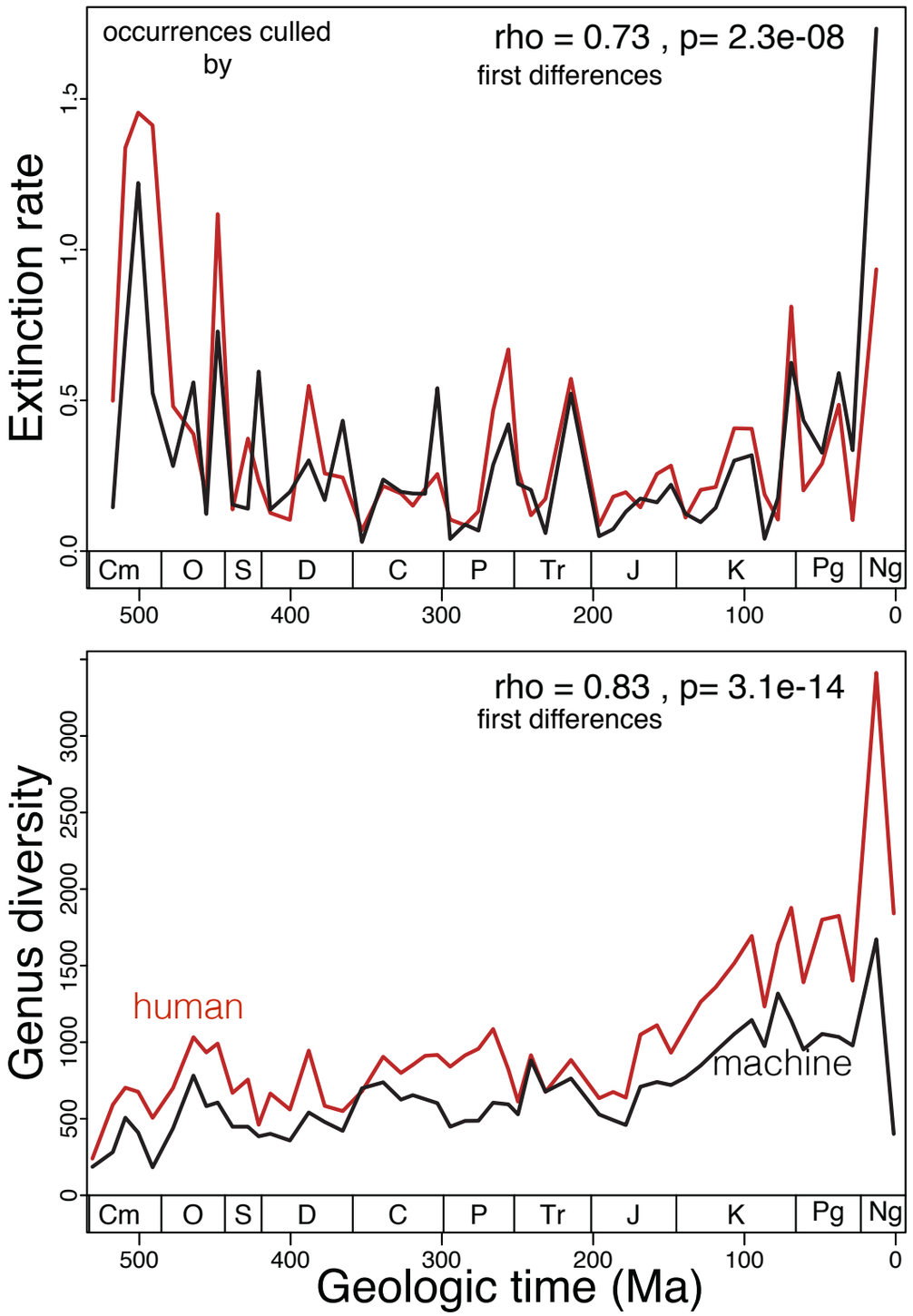
Table Extraction

Table

Age	Formation
Silesian	Tsingyuan Formation

Relational Features

Entity1	Entity2	Feature
Namurian	Tsingyuan Fm.	nn
Silesian	Tsingyuan Fm.	SameRow



PaleoDeepDive

- From a collection of paleontological literature, extracts relations between biological taxa, geological formations, geographic locations, and geological time intervals
- “**PaleoDeepDive performs comparably to humans in several complex data extraction and inference tasks** and generates congruent synthetic results that describe the geological history of taxonomic diversity and genus-level rates of origination and extinction.”
  - "A Machine Reading System for Assembling Synthetic Paleontological Database" (Peters, Zhang, Livny, Re)
  - http://deepdive.stanford.edu/doc/paleo.htm
- Also shows that the quality of the data inferences **systematically improves as information is added.**

We are here,
but
much of the
data are over
there

content owners/providers

WILEY
Publishers Since 1807

USGS
science for a changing world

ELSEVIER

CAMBRIDGE
UNIVERSITY PRESS

Three major hurdles:

TDM user

- Access to documents
- Processing power
- Dependability/repeatability

content owners/providers

WILEY
Publishers Since 1807

USGS
science for a changing world

ELSEVIER

CAMBRIDGE
UNIVERSITY PRESS

A shift in project ambitions…

Let’s build a “smart” library of TDM products!

Three Infrastructure Challenges

- Access to documents
    - Legal and responsible access to scientific literature
- Processing power
    - Need resources, automation, and flexibility
        - This framework should be useful for non-DeepDive applications as well!
- Dependability/Repeatability
    - Track the source of every word/sentence provided to an enduser, and always provide links back to the original content owner.

controlled/authorized access

CENTER FOR HIGH THROUGHPUT COMPUTING

large-scale ***processing jobs***
(using encrypted file system)

content owners/providers

WILEY Publishers Since 1807

USGS science for a changing world

ELSEVIER

CAMBRIDGE UNIVERSITY PRESS

controlled document fetching (key/rate-based)

Website/API

GitHub App

TDM user

DeepDiveSubmit
TDM-ready library, supporting databases

encrypted PDFs for processing,

Database

Secure@SWAMP
original PDF document storage
highly restricted access

controlled/authorized access

CENTER FOR HIGH THROUGHPUT COMPUTING

large-scale ***processing jobs***
(using encrypted file system)

content owners/providers

WILEY Publishers Since 1807

USGS science for a changing world

ELSEVIER

CAMBRIDGE UNIVERSITY PRESS

controlled document fetching (key/rate-based)

Website/API

TDM user

GitHub App

DeepDiveSubmit
TDM-ready library, supporting databases

encrypted PDFs for processing,

Database

Secure@SWAMP
original PDF document storage
highly restricted access

Infrastructure Challenge 1: Legal and Responsible Access to Documents

- Working with the UW Library to draft and sign contracts with large publishers

- Strive to be “Good citizens”

  - Limit ourselves to an agreed-upon fetching rate

  - Providing feedback to publishers, as inconsistent data and system hiccups are discovered

  - Never provide the PDFs themselves to endusers! Only derived products and links back to the provider.

Schedule 1

GeoDeepDive Project

The purpose of the GeoDeepDive Project is to cognitively read in a visual and contextual sense documents in order to recognize the meaning of words, phrases, numbers, and images. This is a machine learning approach such that as the data grow in volume and feedback is gathered from humans, etc., the quality of inference across the entire corpus can improve. Thus the text and data mining activities for this project will also include the ongoing re-reading and re-analysis of content as new rules, signals and dictionaries are developed. The goal is to leverage and expose data in those published works that can, in aggregate, enable new science. The project is interested in attribution, citation, and various metrics for data volume, data quality, and data utility, thus also returning value to publishers.

The Authorized Users for the GeoDeepDive Project t are those faculty, staff, students, and researchers specifically identified as associated with the project, regardless of institutional affiliation, along with the necessary technical support staff.

controlled/authorized access

CENTER FOR HIGH THROUGHPUT COMPUTING

large-scale ***processing jobs***
(using encrypted file system)

content owners/providers

WILEY Publishers Since 1807

USGS science for a changing world

ELSEVIER

CAMBRIDGE UNIVERSITY PRESS

controlled document fetching (key/rate-based)

Website/API

GitHub App

TDM user

DeepDiveSubmit
TDM-ready library, supporting databases

encrypted PDFs for processing,

Database

Secure@SWAMP
original PDF document storage
highly restricted access

CENTER FOR HIGH THROUGHPUT COMPUTING

large-scale ***processing jobs***
(using encrypted file system)

controlled/authorized access

content owners/providers

WILEY Publishers Since 1807

USGS science for a changing world

ELSEVIER

CAMBRIDGE UNIVERSITY PRESS

controlled document fetching (key/rate-based)

Website/API

GitHub App

TDM user

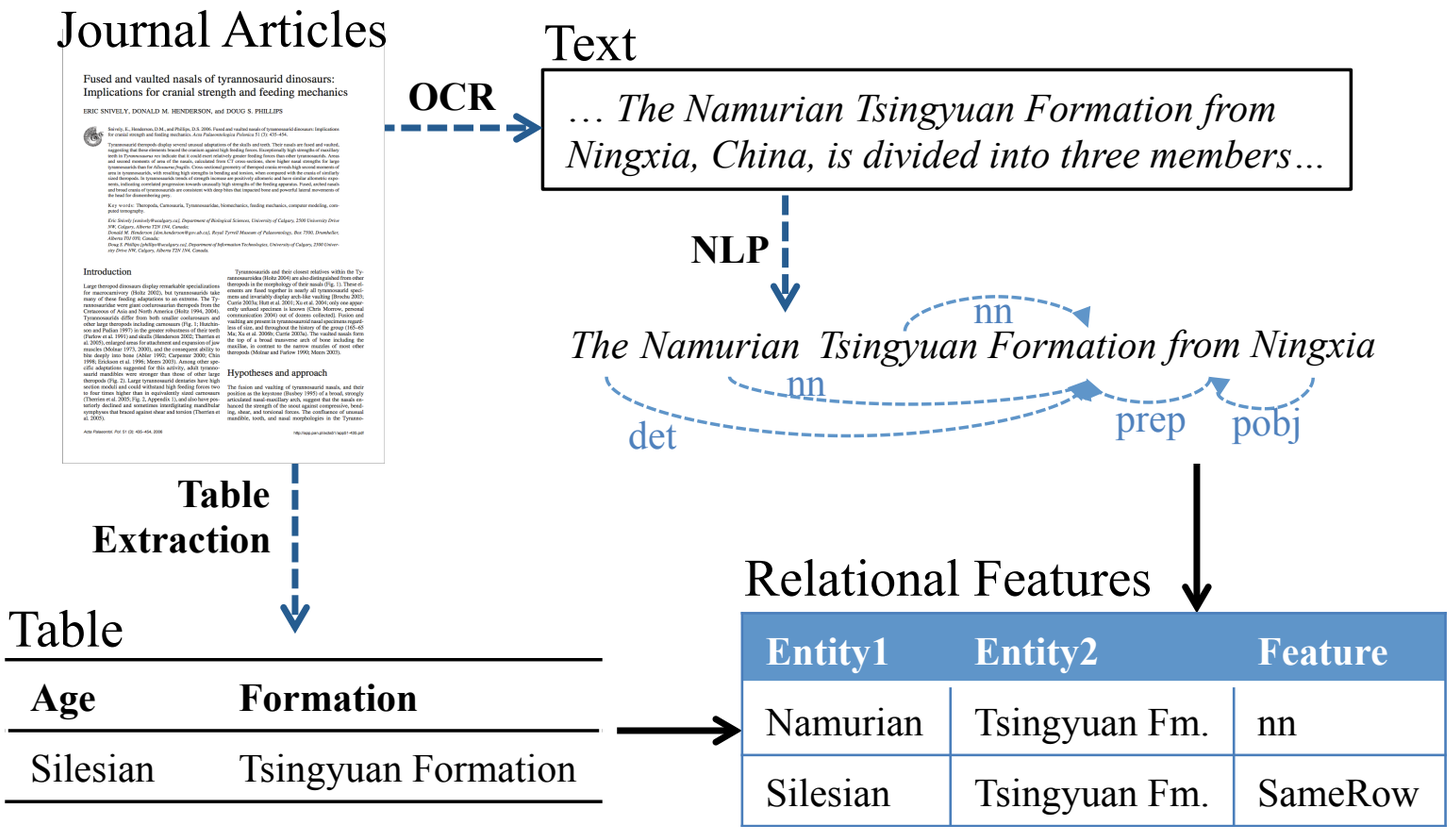
DeepDiveSubmit
TDM-ready library, supporting databases

encrypted PDFs for processing,

Database

Secure@SWAMP
original PDF document storage
highly restricted access

Infrastructure Challenge 2: Processing Power



- 1.2 million articles (+50,000 per week), 6 processing types
- Fairly small/short analysis jobs (3-5 minutes on average for OCR jobs, slightly longer for NLP)
- **High throughput computing is exactly what we need!**
- Use HTCondor and the UW CHTC resources for all of this processing work.

The Infrastructure Challenge — Processing

- Specific needs:
  - Automation and organization — 50,000 articles x 6 different processing types = Potential management nightmare
    - Database makes it easy to ID articles that need processing, keep track of products.
  - Security
  - Flexibility — New tools and document sources should be easy to add to the pipeline
    - New documents are easy — if there's an entry in the database, they'll get processed

The Infrastructure Challenge — Processing

- Specific needs:
  - Automation and organization — processing types = Potential ma
    - Database makes it easy to k track of products.
  - Security
  - Flexibility — New tools and docu to the pipeline
    - New documents are easy — they'll get processed

- Provided by HTCondor!
  - DAGMan w/ postscript to help stay organized (rescue files, dag-level-throttling)
  - Encrypted filesystem ensures PDFs won't be left exposed on the execute nodes

Throughput Statistics

- With a fetch rate of 50,000 articles/week and 6 current processing types, “steady state” requires ~5000 cpu hours per day

- Also have the capability (and resources!) to deploy new processing types across the entire corpus

  - Recently deployed a new version of the Stanford CoreNLP tool to all documents (1 million at the time)

    - Took ~3 weeks to process 1 million documents, while still staying up to date with the other processing types

controlled/authorized access

CENTER FOR HIGH THROUGHPUT COMPUTING

large-scale ***processing jobs***
(using encrypted file system)

content owners/providers

WILEY Publishers Since 1807

USGS science for a changing world

ELSEVIER

CAMBRIDGE UNIVERSITY PRESS

controlled document fetching (key/rate-based)

Website/API

GitHub App

TDM user

DeepDiveSubmit
TDM-ready library, supporting databases

encrypted PDFs for processing,

Database

Secure@SWAMP
original PDF document storage
highly restricted access

controlled/authorized access

CENTER FOR HIGH THROUGHPUT COMPUTING

large-scale ***processing jobs***
(using encrypted file system)

content owners/providers

WILEY Publishers Since 1807

USGS science for a changing world

ELSEVIER

CAMBRIDGE UNIVERSITY PRESS

controlled document fetching (key/rate-based)

Website/API

GitHub App

TDM user

DeepDiveSubmit
TDM-ready library, supporting databases

encrypted PDFs for processing,

Database

Secure@SWAMP
original PDF document storage
highly restricted access

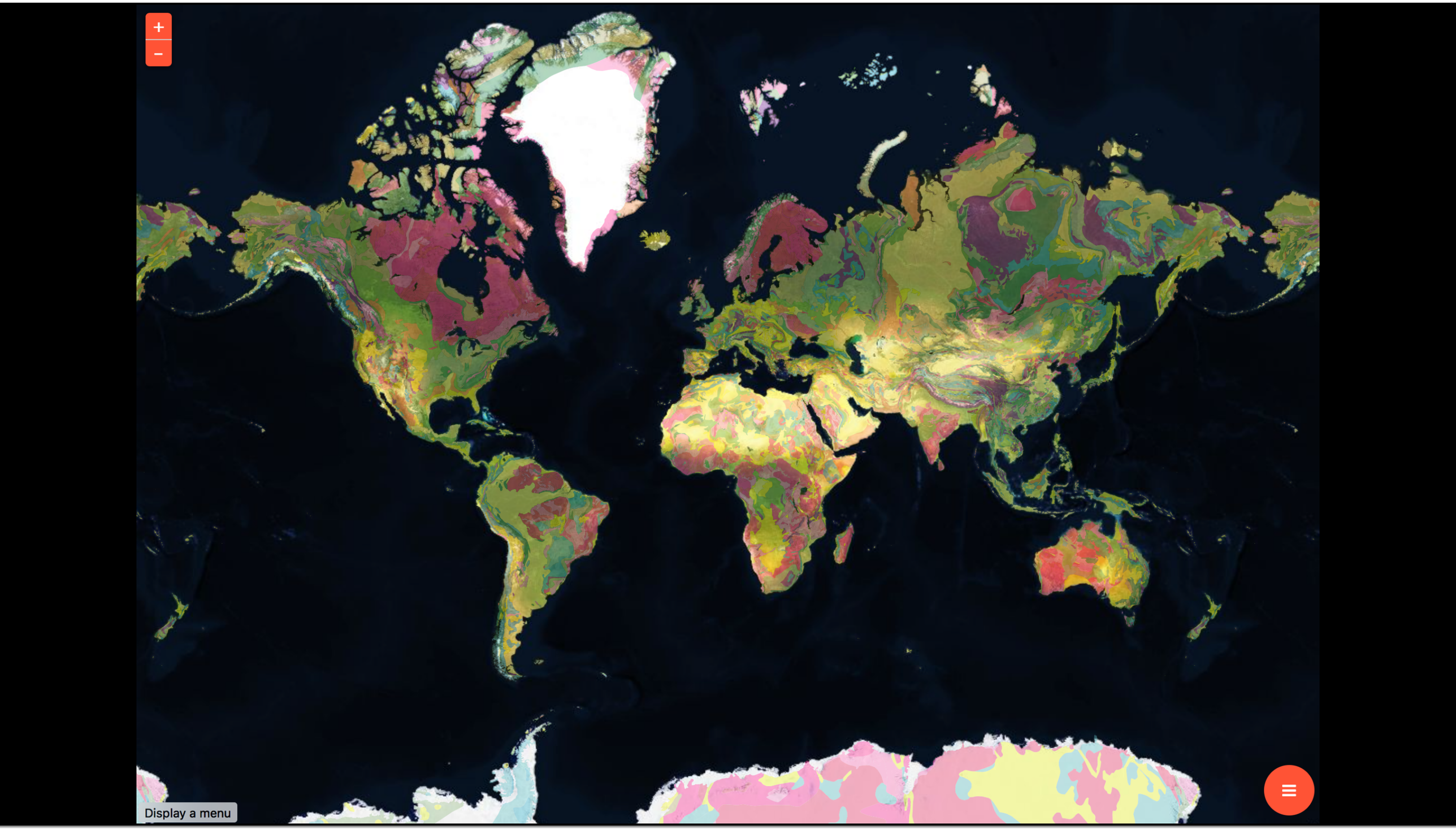
What can be done with all this data?

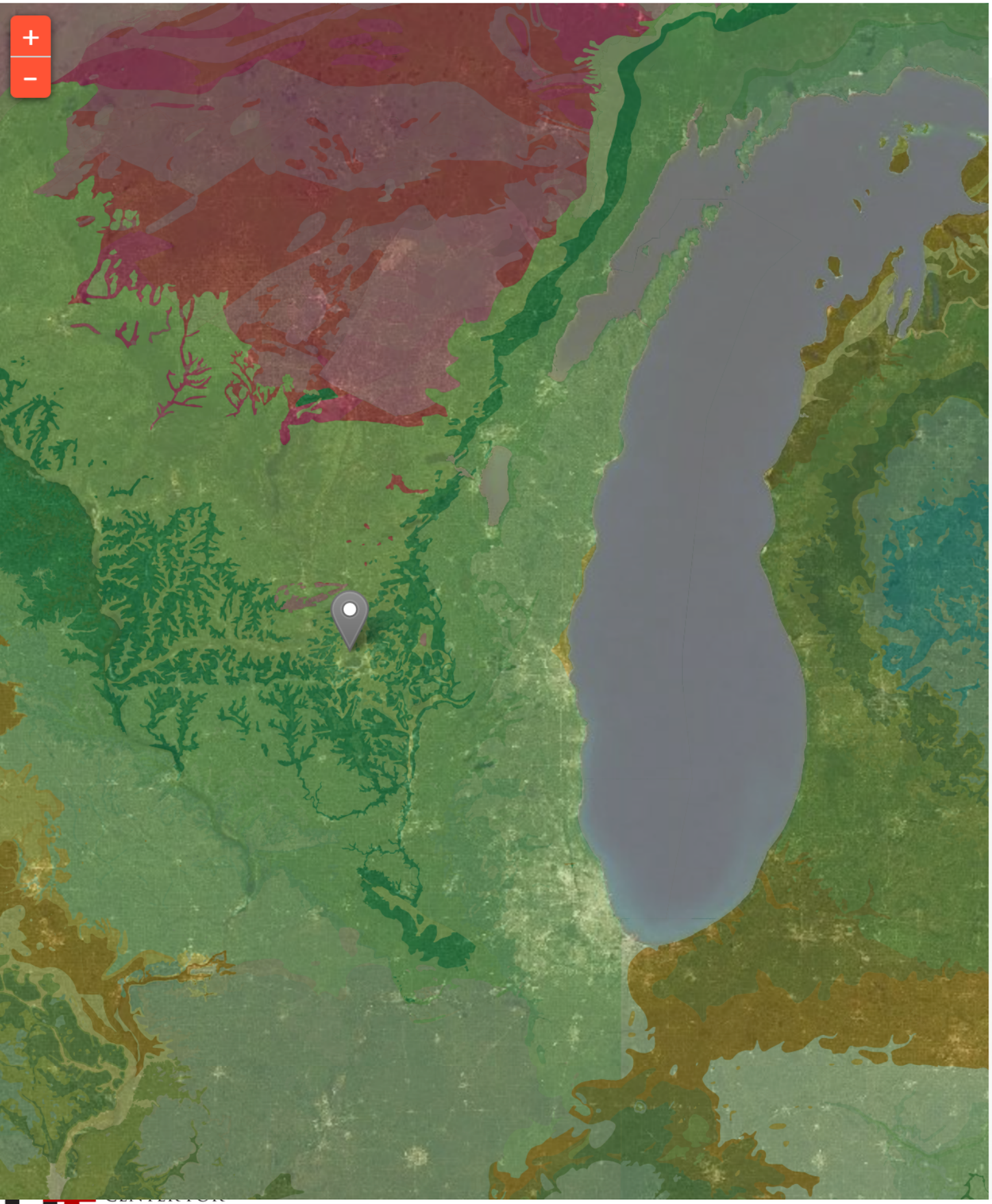
- PaleoDeepDive showed that machine reading can infer facts and build a database comparable to years of human effort

- Even without bringing machine learning into the picture, there's a huge amount of value in the sentence-level data and NLP products!

e.g. Space-time index of the literature

https://macrostrat.org/burwell/





X

43.1491, -89.4507 276 m | 906 ft

Tunnel City Group, Elk Mound Group

Age: Dresbachian - Trempealeauan - (499.95 - 487.175 Ma)

Thickness: 0 - 150m

PBDB Collections: 4

Unit IDs: 6052, 6053, 6104, 6105, 6106, 6117, 6119, 6120, 6126, 6127, 6128, 6131, 6132, 6133, 6141, 6142, 6143, 6144, 6146, 6252

Reference: Macrostrat.org

Professional Paper USGS

Young, H. L., Siegel, D. I., 1992. **Hydrogeology of the Cambrian-Ordovician aquifer system in the northern Midwest, United States, with a section on ground-water quality.**

...----------9. Eau Claire Formation or Bonneterre Formation...

...underlying Eau Claire Formation and its partial equivalent to the southwest, the Bonneterre Formation...

.... Siltstone and shale are fairly common in the upper part of the Eau Claire Formation but less so in its...

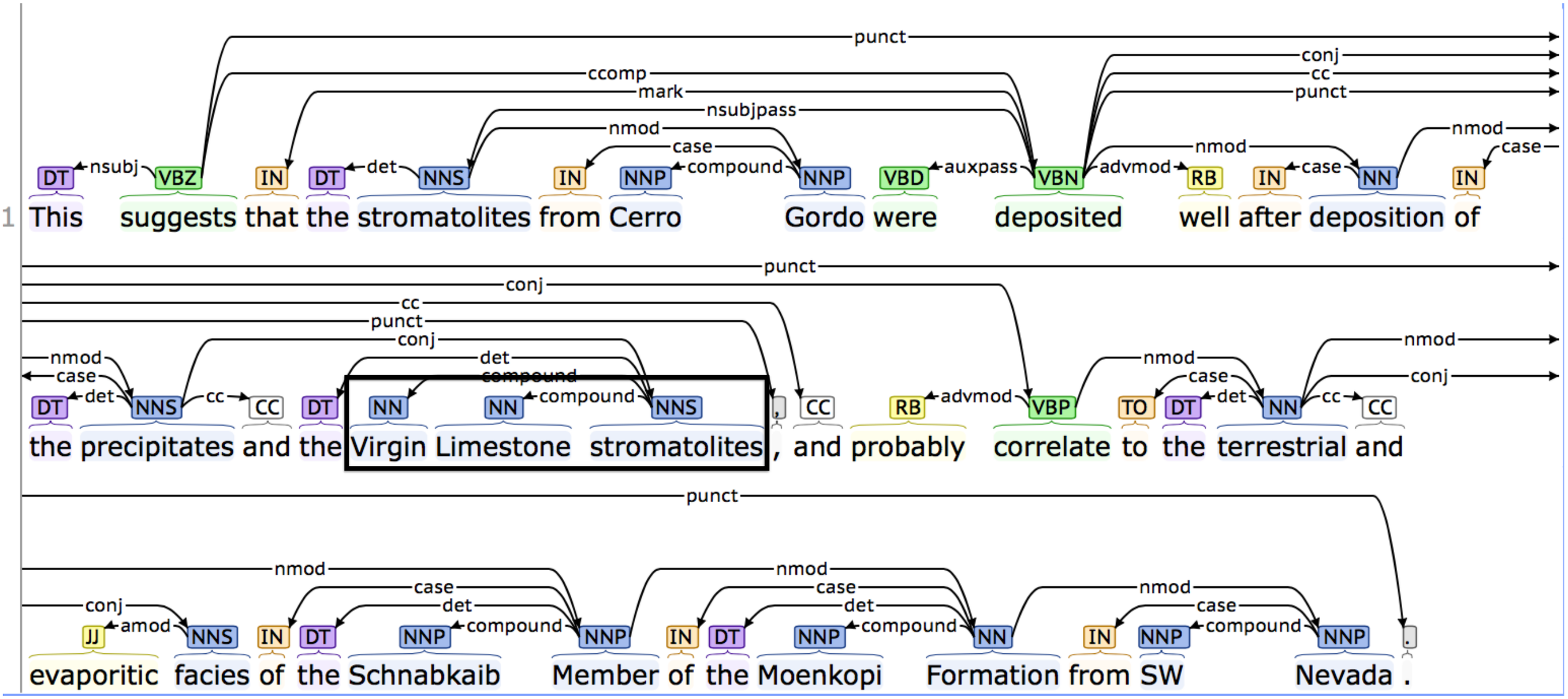
...the Eau Claire Formation in northern Illinois. The aquifer increases greatly in thickness and the...

...Biogenic Depositional shelf St. Lawrence Formation Tunnel City Group Van Oser Member Norwalk...

Bridge, Josiah, 1937. **The correlation of the Upper Cambrian sections of Missouri and Texas with the section in the upper Mississippi Valley.**

e.g. New synthetic results

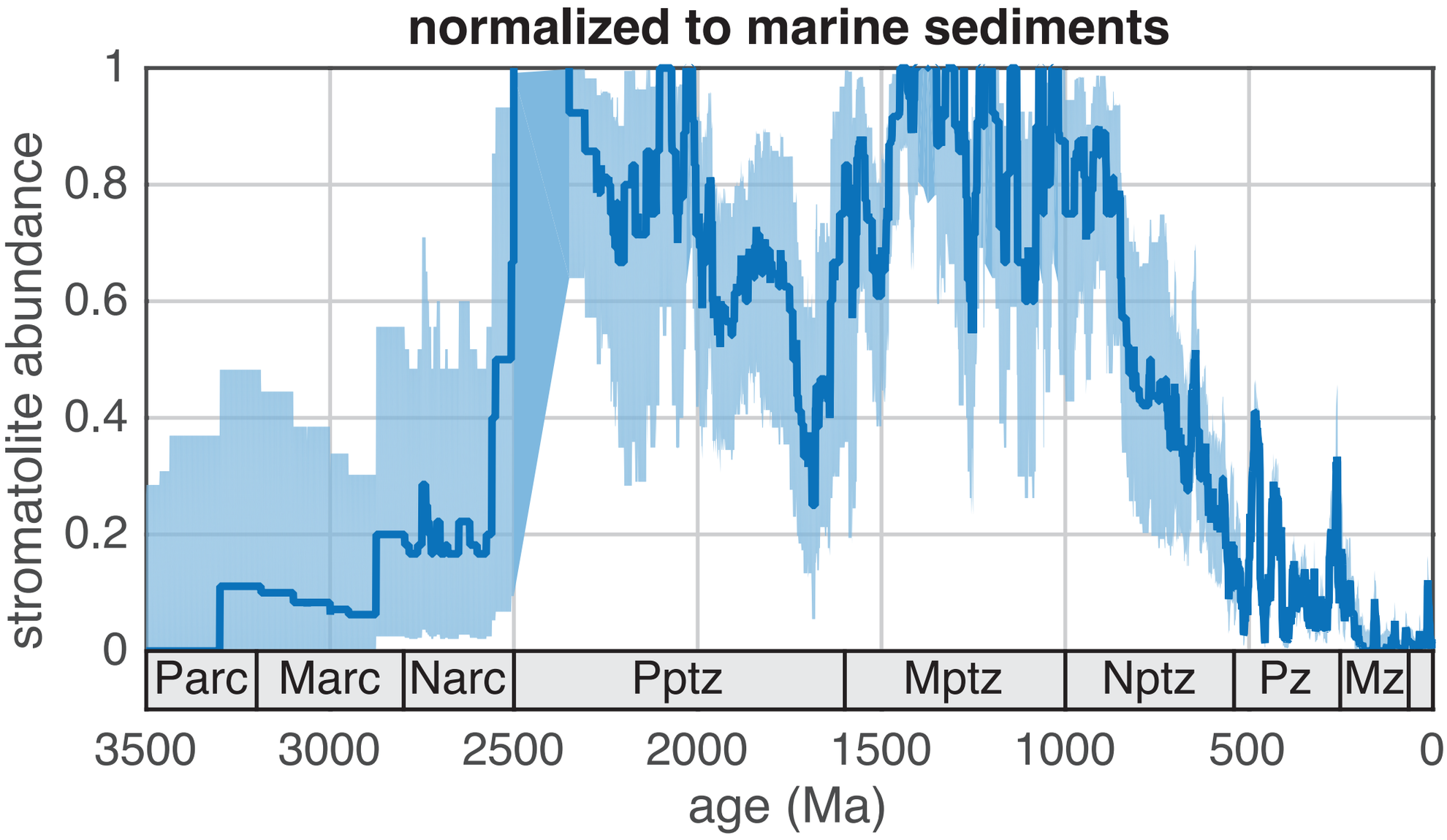
GeoDeepDive + Macrostrat tuple extraction: lots of entities, NLP features link them



Stromatolite prevalence in the geologic record

Julia Wilcots

Jon Husson



Conclusions — Key Infrastructure Features

- Automated document fetching at arbitrary maximum rates determined by content providers (e.g., Elsevier 10K/week/API key)

- Secure document storage; encrypted processing methods to protect content owners/providers

- HTC infrastructure to run core tools (e.g., NLP, OCR, table recognition/ parsing, image analysis), with flexibility and power to add more.

- API layer with basic capacity to identify documents of potential relevance to a project, with initial results returned as (augmented) bibJSON

- Packaging and delivery of analysis-ready sentence data (e.g., PostgreSQL database of NLP results); everything traceable back to specific sources (original URL and locations within documents)

Lots of ways to get involved!

- Identify and help retrieve documents from content owners (e.g., museum publications series, society publications, open-access content)

- Write TDM applications that can facilitate your work/science and do cool things; *we will help you*!

- Develop tools for parsing/reading documents in your area of work; develop comprehensive dictionaries of terms in your field and make them accessible to us so we can pre-index the literature

- Leverage our APIs in your applications (just let us know, we might help)!

Questions?

- http://www.geodeepdive.org

- iross@cs.wisc.edu